

IMLS GRANT LG 0302-0100-02  
Interim Performance Report 1  
1 October 2002 – 31 March 2003  
Submitted by Principal Investigator Priscilla Caplan

The goals of this project are: 1) to establish a working digital preservation archive for the use of the libraries of the public universities of Florida (the FCLA Digital Archive, or FDA), 2) to identify costs involved with sufficient granularity to support reasonable cost-recovery pricing, and 3) to disseminate tools, procedures and results for the widest national impact.

Towards the first goal, establishing a working digital preservation archive, the following activities have been accomplished within the reporting period:

- A full time formats specialist with all desired qualifications was hired. She started work in mid-January, only a few weeks behind the target start date of January 1.
- A project team consisting of the P.I., the formats specialist, three programmers and a systems administrator was formed. The team has been meeting weekly since January.
- The project team found that the lack of a common vocabulary for referring to archival concepts and processes was a problem. A set of terms and their definitions was drafted for our own working use and is available on the website.
- Alternate preservation strategies have been studied and modeled. The working decision is to use a combination of normalization, mass migration, and migration on request.
- The overall program architecture of the archive, beginning with the ingest function, is being modeled in Poseidon, a Java-based UML (unified modeling language) and CASE tool.
- Data tables for the archive management system, including administrative and preservation metadata, are defined and documented in a data dictionary.
- “Action plans” for preservation treatment are defined for 6 file formats: PDF 1.2, PDF 1.3, PDF 1.4, plain text, TIFF 5.0 and TIFF 6.0.
- A policy guide defining respective roles and responsibilities of the libraries and the FDA was drafted. A model contract between libraries and the FDA has been submitted to legal counsel for review.
- Hardware and software needs were defined. A workstation for the formats specialist was purchased. A standalone machine specifically for archival processing has been specified for purchase.

- Programming is expected to start in May, a month behind schedule.

Towards the second goal, identifying costs, no steps have been taken beyond detailed monthly time accounting by all project participants. Time is reported in these categories: R&D, software development, storage management, publicity/promotion, production, and administration.

It is a little premature for the third goal, dissemination, but some early steps have been taken.

- The project website is available at [www.fcla.edu/digitalArchive/](http://www.fcla.edu/digitalArchive/). Format-specific action plans and other documents are posted as they are reviewed and accepted by the project team. The project's Outcomes Logic Model is available from this site at [www.fcla.edu/digitalArchive/pdfs/DigitalArchiveOutcomesModel.pdf](http://www.fcla.edu/digitalArchive/pdfs/DigitalArchiveOutcomesModel.pdf).
- Contacts have been established between project staff and technical staff at many other organizations involved in digital archiving.
- An article about our experiences so far will be published in a special issue of VINE on digital archiving to appear in early 2004.

As of the end of the reporting period, the project is proceeding according to the schedule of completion in the project plan, and the archive is on target to being ingesting materials on July 1, 2003. However, the project team has found that many issues are more complicated than first appeared, and a great deal of time must be spent on high level design and repeated modeling of events.

For example, it was initially assumed that participating libraries would send Submission Information Packages (SIPs) to be archived according to the OAIS model, and that these packages would contain all of the content information to be preserved along with METS-formatted metadata that we call the "SIP descriptor." In reality, however, the packages already queued for archiving are incomplete, as they rarely contain all of the XML schema referenced by the METS schema. The archive ingest function, then, must try to obtain the referenced schema and add these to the package, therefore creating a new SIP and invalidating the original SIP descriptor.

Similarly, the need to archive XML schema has raised the issue of whether Archival Information Packages (AIPs) should be complete in themselves or whether they can reference objects archived in other AIPs. For example, could a Dublin Core schema definition be archived once and pointed to by multiple AIPs, or must the schema be archived with each AIP?

Working through a long chain of issues like these has slowed the design process and demonstrated a need for careful modeling before programming begins.